Unpacking the Unpredictable

From Machine Learning to Generative AI for Detecting Crises in Presidential Democracies

Dr. Bastián González-Bustamante
Associate Professor, Universidad Diego Portales
Lecturer and Postdoctoral Researcher, Leiden University

Presentation at the PoliCICS Workshop, Universidad del Desarrollo
October 9, 2025

From Machine Learning...

Machine Learning and Political Events

DPhil Methodological Paper

Report

Machine Learning and Political Events: Application of a Semi-supervised Approach to Produce a Dataset on Presidential Cabinets

Social Science Computer Review
2025, Vol. 0(0) 1–14

DOI: 10.1177/08944393251315917
journals.sagepub.com/home/ssc
Sage

Bastián González-Bustamante[1,2]

Abstract
This paper describes the creation of a novel dataset on ministerial turnover and resignation calls in 12 presidential cabinets in Latin America from the mid-1970s to the early 2020s. The indicators on resignation calls and reallocations of cabinet members are entirely novel. Both constitute a relevant empirical contribution not only to the study of political dynamics in presidential systems and cabinet politics but also to public opinion and public policy topics. We focus on the creation of the dataset using optical recognition algorithms on press report archives together with machine learning models. The models permitted the training of ensemble semi-supervised classifiers over a period of almost 50 years. Subsequently, we provide a number of measurement validity checks to cross-validate the dataset by comparing it with similar existing data and an exploratory analysis.

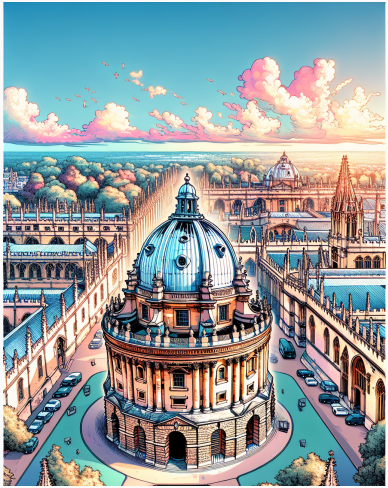
Keywords
comparative politics, executives, cabinets, computational social science, machine learning

Introduction

The scarcity of specific information on political events in some regions can be partly explained by the fact that, although techniques now exist for the automatic production of event data in political science, manual coding remains a common practice. Manual data gathering is costly in terms of time and resources and cannot ensure the reproducibility of the process (Eshima et al., 2021; Radford, 2019). However, since the mid-1990s, there have been growing efforts to use automated coding to produce datasets of events and phenomena relevant to political science and public affairs (e.g. Boschee et al., 2015; Schrodt, 1998, 2011; Schrodt & Brackle, 2013).

[1]Institute of Public Administration, Leiden University, The Netherlands
[2]School of Public Administration, Universidad Diego Portales, Chile

Corresponding Author:
Bastián González-Bustamante, Institute of Public Administration, Faculty of Governance and Global Affairs, Leiden University, Wijnhaven, Turfmarkt 99, The Hague 2511 DP, The Netherlands.
Emails: b.a.gonzalez.bustamante@fgga.leidenuniv.nl; bastian.gonzalez.b@mail.udp.cl



Artwork by DALL·E 3 model

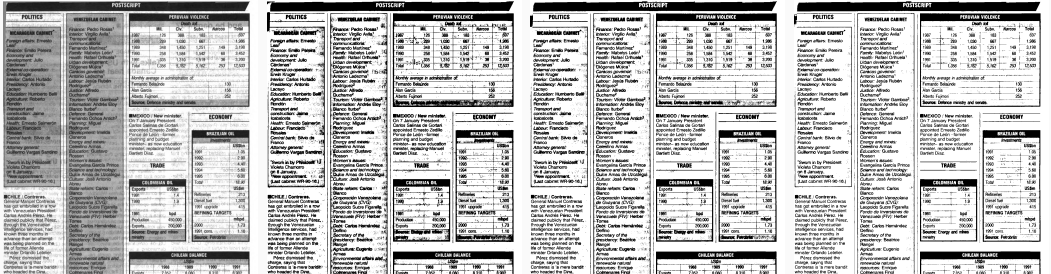
Cases and Data Collection

I combined the Tesseract Optical Character Recognition (**OCR**) with different **semi-supervised machine learning models** to distinguish between resignation calls and other mentions of cabinet members in the press. This was applied to **46 years** of **Latin American Weekly Report (LAWR) archives**.

Detail of the Batches of LAWR Archives

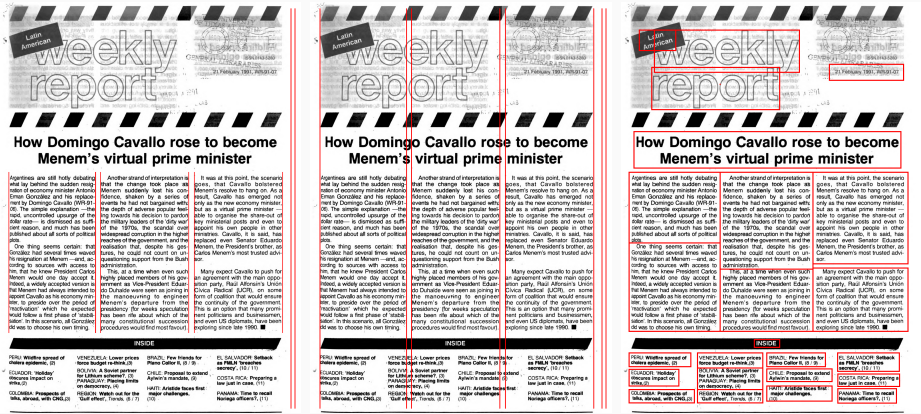
Batch	Archives	Years	Format	Volume
Batch-01	LAWR-1975 to LAWR-1979 Including 1997	5	CSV UTF-8	9.56 MB
Batch-02	LAWR-1980 to LAWR-1998 Excluding 1997 Including the first half of 2003	18.5	PNG images	38.5 GB
Batch-03	LAWR-1999 to LAWR-2021 Excluding the first half of 2003	22.5	PDF files	341 MB

Binarisation Techniques for Enhancing OCR



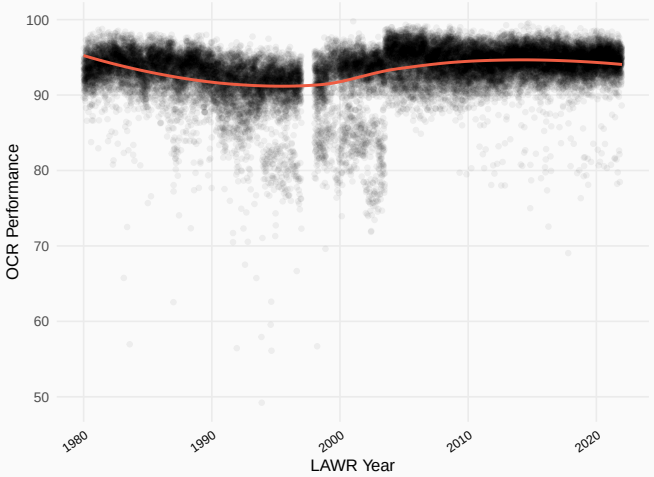
Note. The images correspond to a LAWR page before and after binarisation techniques.

Tesseract Optical Character Recognition



Tab-Stop Lines — Column Layout — Segmented Blocks

Performance and Accuracy of the Tesseract OCR Algorithm



Accuracy is based on the proportion of text that was correctly identified.

This calculation was performed after extracting stop words and carried out with dictionaries trained with **Long Short-Term Memory (LSTM) models** used by Tesseract.

Semi-Supervised Approach

There are two significant **drawbacks** to training **supervised algorithms**:

1. This would have required a major amount of labelled text data **(time and resources)**
2. The accuracy of the algorithms is expected to decrease out-of-the-sample and over time **(batches of archives cover almost 50 years)**

Therefore, I preferred a **semi-supervised approach** that could maintain a **high accuracy level over time with a small amount of coded data** for training.

1. I drew a subsample of 1,000 LAWR press releases
2. A team of human coders classified all these observations (**resignation calls**)
3. Labelling was conducted to generate ground-truth labels to cross-validate
4. Semi-supervised models using knowledge-based seed words

Semi-Supervised Approach

Pre-processing

80% of the subsample for training

20% for validation

100 tokens per news item *

SMART stop words **

Customised list ***

Stemming algorithm

* until 500 for robustness checks

** Snowball for robustness checks

*** based on LAWR numbering

Models

Naïve Bayes (NB)

Support Vector Machine (SVM) *

Radial Basis Function (RBF) SVM **

Random Forest with 100 trees (RF100)

Random Forest with 500 trees (RF500)

Extreme Gradient Boosting (XGBoost)

* Kernel-based Linear

** Kernel-based Gaussian

Labelling Process for Benchmarking Supervised Models

Coder	Labels	$M(T)$ Label	$\sum$ Time	Consensus
1	966	39s	10h 25m	94.05
2	718	1m 15s	15h 0m	94.84
3	704	38s	7h 29m	94.82
4	318	1m 4s	5h 38m	95.47
5	263	17s	1h 14m	99.26
6	101	1m 17s	2h 9m	88.50

Note. Each observation was hand-coded by three different coders. 2.3% of observations were revised and relabelled during an iteration step.

Goodness-of-Prediction of the Semi-Supervised Models

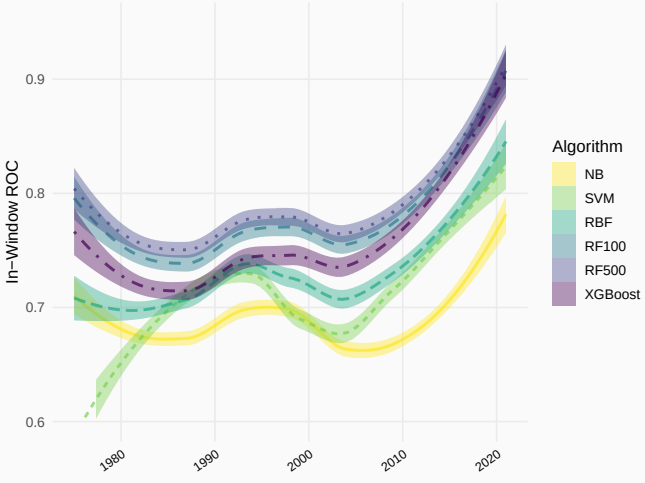
Model	Knowledge-Based Seed Words				
	Accuracy	Precision	Recall	F1-score	ROC
Naïve Bayes	0.713	0.734	0.948	0.828	0.638
Kernel Linear SVM	0.698	0.741	0.900	0.812	0.614
Kernel Gaussian RBF SVM	0.726	0.728	0.995	0.840	0.647
Random Forest (100 trees)	0.743	0.742	0.991	0.849	0.671
Random Forest (500 trees)	0.743	0.740	0.997	0.849	0.682
XGBoost	0.726	0.762	0.905	0.827	0.660

Note. The evaluation was performed with resampling methods using the testing set and 10-fold cross-validation.

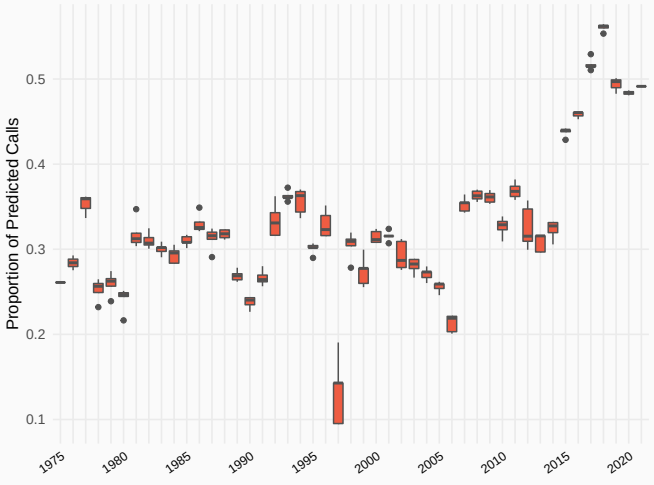
Goodness-of-Prediction of Semi-Supervised Models (1975-2021)

I trained semi-supervised models using a **five-year fixed rolling window** from 1975 to 2021 to train algorithms and predict resignation calls.

Random Forest classifiers with 500 trees 10-fold cross-validated.



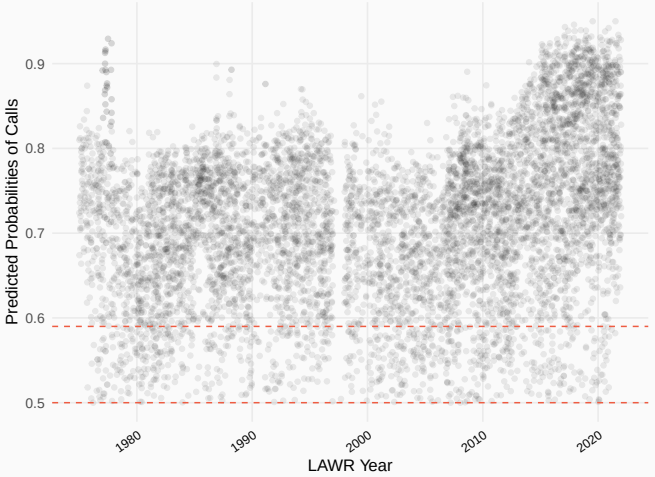
Predicted Ministerial Resignation Calls over Cabinet Press Mentions



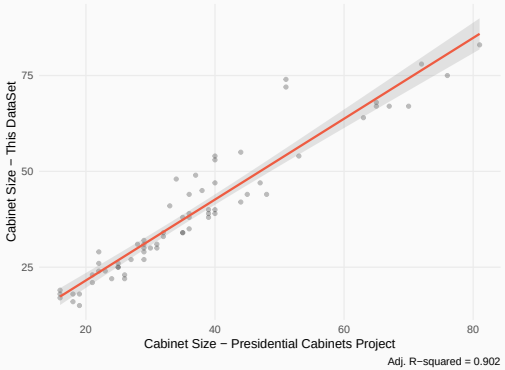
Proportion of predicted resignation calls over mentions of cabinet members in press reports, were carried out with an **ensemble semi-supervised RF500 algorithm**.

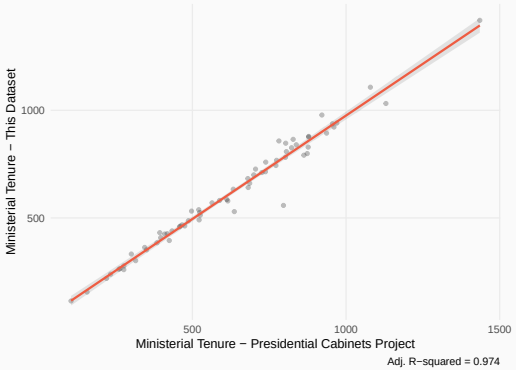
Human-in-the-Loop

I performed an augmented artificial intelligence process based on a **human review** of the predicted observations with low confidence values to improve the data's quality.



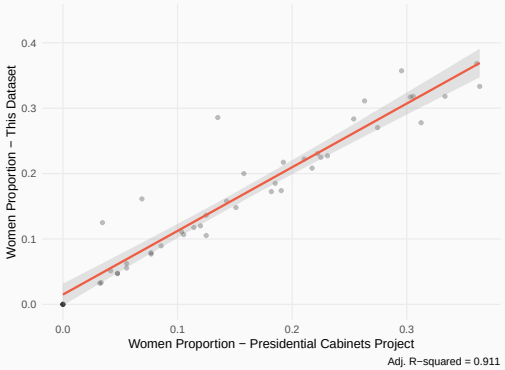
Convergent Validation against Similar Datasets

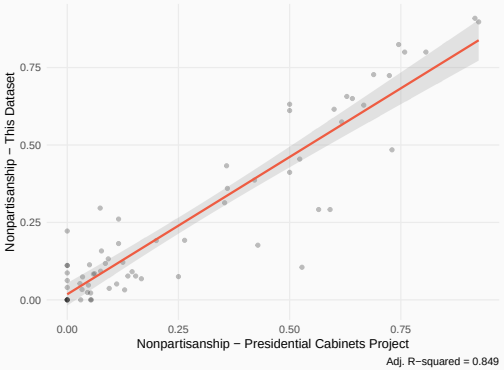




Note. Concurrent presidential terms were used for comparisons with the Presidential Cabinet Project.

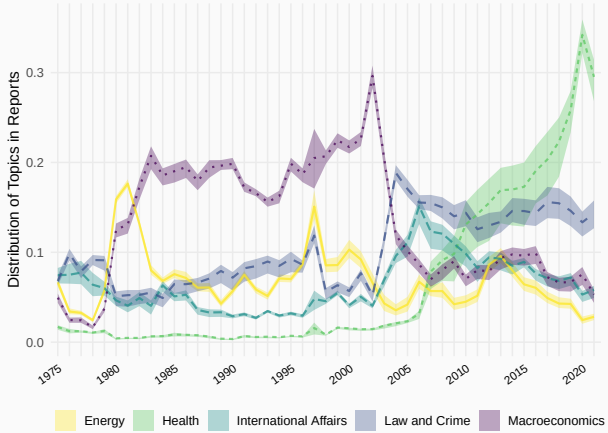
Convergent Validation against Similar Datasets





Note. Concurrent presidential terms were used for comparisons with the Presidential Cabinet Project.

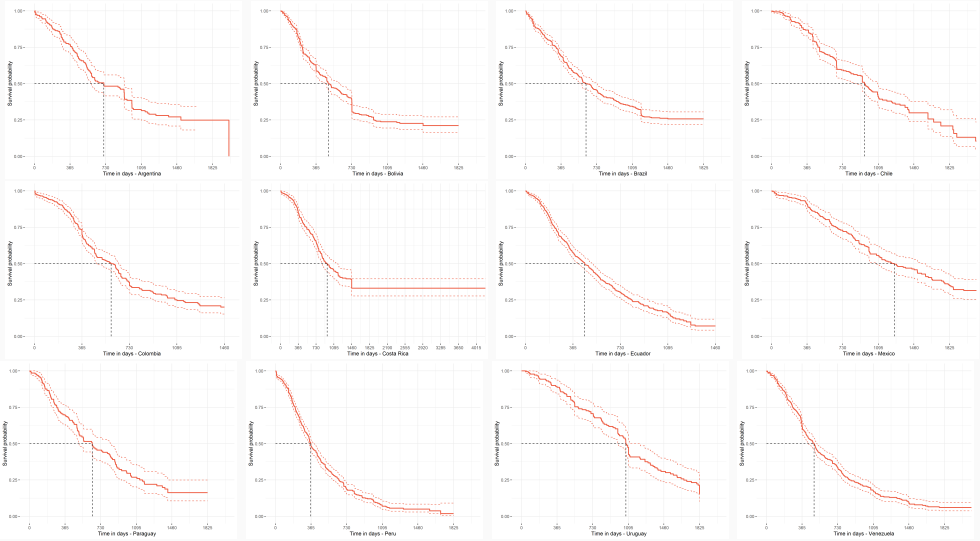
Content Validation: Major Agenda Topics in Latin America



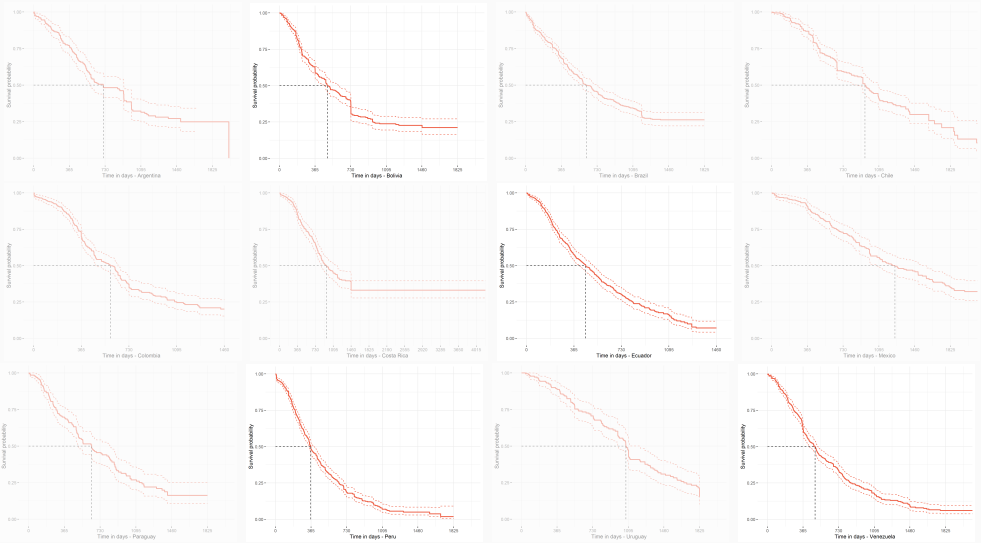
I used **Keyword Assisted Topic Modelling** (**keyATM**) and a dictionary based on Comprative Agendas Project (CAP).

It is a semi-supervised Bayesian application that allows me to obtain ϑ (distribution of topics in the documents and CIs).

Kaplan-Meier Survival Estimations in 12 Presidential Cabinets

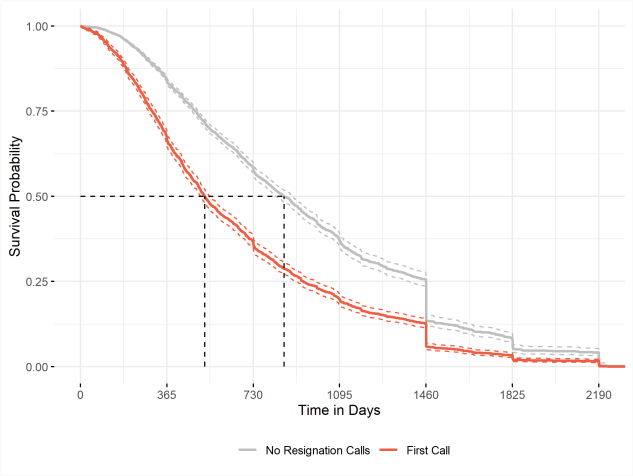


Kaplan-Meier Survival Estimations in 12 Presidential Cabinets



Kaplan-Meier Survival Estimation

$F(t) = 1 - S(t) = Pr(T \leq t)$ can be used to obtain the incidence of ministerial exits during T by considering $S(t)$ as a survival function, if raw exit is considered an event in the **absence of competing risks**.



... to Generative AI

Enlace-Inserción UDP 2025

Unpacking the Unpredictable: Cabinet Politics in Presidential Democracies

Examining how NLP and LLMs can shed light on cabinet responses to stochastic events in Latin American democracies

Dr. Bastián González-Bustamante, Principal Investigator

Enlace-UDP Design

Project Overview

Artwork by DALL·E 3 model

Focus. Investigating the impact of stochastic events on cabinet stability in presidential systems.

Scope. 12 Latin American democracies, mid-1970s to early 2020s.

Methodology. Utilising machine learning and AI techniques, including LLMs.

Introducing LLMs to the Field

Traditional text analysis. Traditional approaches to analysing text data in social sciences often relied on manual coding, dictionaries, or topic modelling, which can be time-consuming and resource-intensive.

The power of LLMs. LLMs, with their zero- and few-shot learning capabilities, offer a powerful alternative, enabling researchers to process and analyse vast amounts of text data efficiently.

Transforming research. This shift towards AI-powered analysis has the potential to transform computational social science research.

OpenAI's GPTs: A Closer Look

GPT and research

Researchers have increasingly utilised OpenAI's GPT models through their **API for various tasks**, including text annotation and classification.

Advantages of GPT models

GPT models offer several advantages, including **ease of use**, access to substantial computational resources, and **impressive performance** in many tasks.

Concerns and limitations

However, concerns remain regarding **reproducibility**, data privacy, and the potential for bias in these proprietary models.

The Rise of Open-Source LLMs

Addressing the limitations

Open-source LLMs have emerged as a response to concerns surrounding proprietary models, emphasising transparency, **reproducibility**, and community-driven development.

Performance and potential

Recent studies have shown that open-source LLMs can match or **even surpass** the performance of proprietary models in specific text annotation tasks, particularly in zero-shot classification.

Cost-effective alternatives

Some open-source "Small Language Models" (SLMs) offer comparable performance with significantly **reduced computational requirements** and costs.

Case Selection Strategy

Brazil

South America's largest democracy, featuring a complex federal system and a diverse political landscape.

Venezuela

Represents significant institutional changes and varying levels of democratic stability over the study period.

Costa Rica

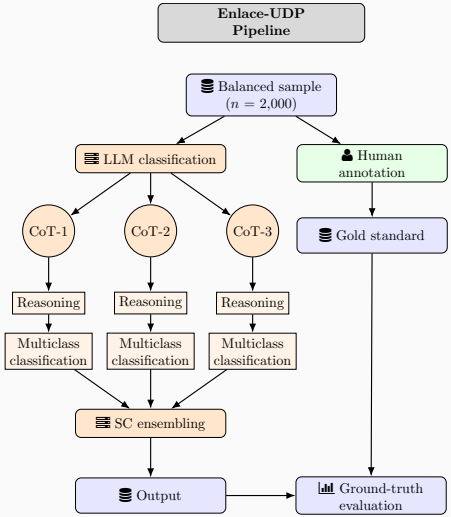
Known for consistent democratic governance and a stable institutional framework.

Mexico

Features transition from dominant-party system to competitive democracy.

Project Pipeline

- 500 media reports per country.
- 3 human coders per report.
- **Gold standard** creation and measurement validity assessment.
- We may slightly increase the sample and go for some **fine-tuned models**.



Benchmarking LLMs for Annotation

Benchmarking LLMs for Annotation

Measuring Complexity and Reproducibility: A Comprehensive Benchmark of LLMs for Multilingual Policy Agenda Topic Annotation*

Bastián González-Bustamante†
Leiden University

June 3, 2025

Abstract

Although Large Language Models (LLMs) have rapidly expanded the text-as-data toolkit available to social scientists, their performance remains highly sensitive to task complexity, prompting, language and model parameters. In order to assess these limits, we conduct a cross-lingual benchmark of 79 contemporary LLMs (e.g., OpenAI's GPTs and o-series, Claude models, xAI's Grok, Meta's Llama, Alibaba's Qwen, Mistral models, among others) on a demanding zero-shot classification task: labelling one of the 21 major topics in the Comparative Agendas Project codebook to bills and Acts drawn from Danish, Dutch, English, French, Hungarian, Italian, Portuguese and Spanish corpora. We then conducted a meta-analysis that shows reasoning models lift F1-scores by around nine percentage points on average. In comparison, reproducible models under deterministic deployment conditions incur a cost of about eight percentage points in F1-score. In addition, preliminary and ongoing fine-tuning experiments suggest that fine-tuned transformers may surpass the best zero-shot LLM classification performance. These findings quantify the trade-off between performance and reproducibility, highlighting chain-of-though reasoning as

*This version was prepared for delivery at the Annual Conference of the Comparative Agendas Project (CAP), University of Konstanz, Konstanz, Germany, June 11–13, 2025.

†Post-doctoral Researcher in Computational Social Science and Lecturer in Governance and Development, Institute of Public Administration, Faculty of Governance and Global Affairs, Leiden University, Netherlands. Wijnhaven, Turfmarkt 99, The Hague 2511 DP, Netherlands, b.a.gonzalez.bustamante@fgga.leidenuniv.nl, https://bgonzalezbustamante.com, ORCID iD https://orcid.org/0000-0003-1510-6820.

Charting Reproducibility and Performance: LLMs in Multilingual Toxic Speech Detection*

Bastián González-Bustamante†
Leiden University

June 10, 2025

Abstract

Large Language Models (LLMs) are increasingly central to analysing and mitigating incivility and toxicity in online communication, however, their comparative strengths vary by language coverage, model openness, and other factors. Drawing on the Multilingual Text Detoxification (TextDetox) corpus, which spans seven languages (i.e., Arabic, Chinese, English, German, Hindi, Russian, and Spanish), this paper benchmarks 807 model-language pairs and pools goodness-of-prediction indicators in a meta-analysis. The evaluation includes OpenAI's GPTs, o-series, Claude models, xAI's Grok, Llama checkpoints, Alibaba Qwen-series, Mistral models, among others. Three patterns emerge from our analysis. First, high-resource languages (i.e., English, German, and Spanish) enjoy, on average, a 7.7-point F1-score advantage over lower-resource counterparts (i.e., Arabic, Chinese, Hindi, and Russian). Second, proprietary models top the leaderboard in low-resource languages, however, the openness penalty is minor and statistically inconclusive. Conversely, open-source models tend to equal closed models in high-resource languages. Third, reasoning models and chain-of-thought (CoT) neither help nor harm this binary classification task, whereas compact models (with 7B parameters or fewer) trail a larger baseline by roughly 9.8 points. Ancillary findings derived from temperature

*This version was prepared for delivery at the LLM Pre-Conference Workshop, European Political Science Association (EPSA), Universidad Carlos III de Madrid, Madrid, Spain, June 25, 2025. An early version of this work was presented at the 7th Annual COMPTEXT Conference, University of Vienna, Austria, April 24-26, 2025. I am grateful to Chico Camargo for his helpful comments.

†Post-doctoral Researcher in Computational Social Science and Lecturer in Governance and Development, Institute of Public Administration, Faculty of Governance and Global Affairs, Leiden University, Netherlands. Wijnhaven, Turfmarkt 99, The Hague 2511 DP, Netherlands, b.a.gonzalez.bustamante@fgga.leidenuniv.nl, https://bgonzalezbustamante.com, ORCID iD https://orcid.org/0000-0003-1510-6820.

Ground-Truth Policy Agenda

Denmark	NLD	UK	France
15 101 bills 1953–2016	4 684 bills 1981–2009	6 169 Acts 1911–2015	3 069 laws 1979–2013

Hungary	Italy	Brazil	Spain
8 220 bills 1990–2022	4 554 laws 1983–2013	2 449 laws 2003–2014	2 256 laws-decrees 1980–2018

Note. We split the samples in a proportion of 70/15/15 (stratified by major topic) for training, validation, and testing for future fine-tuning jobs. The samples correspond to ground-truth data of the Comparative Agendas Project.

Ground-Truth Toxicity

Arabic
5000 tweets manually annotated

Chinese
5000 messages for toxic detection

English
5000 Wikipedia comments

German
5000 Twitter and Facebook comments

Hindi
5000 Twitter and Facebook comments

Russian
5000 comments on social network OK

Spanish
5000 messages for toxic detection

Note. We split the samples in a proportion of 70/15/15 for training, validation, and testing in case of fine-tuning jobs. The samples correspond to ground-truth data prepared for CLEF TextDetox (2024).

LLMs Zero-Shot Classification

79/109 LLMs

Multiclass/Binary

Run **620/807 times** under different conditions (e.g., parameters, API/local, temperature, datasets/language)

* GPT-5 and OSS (August 2025).

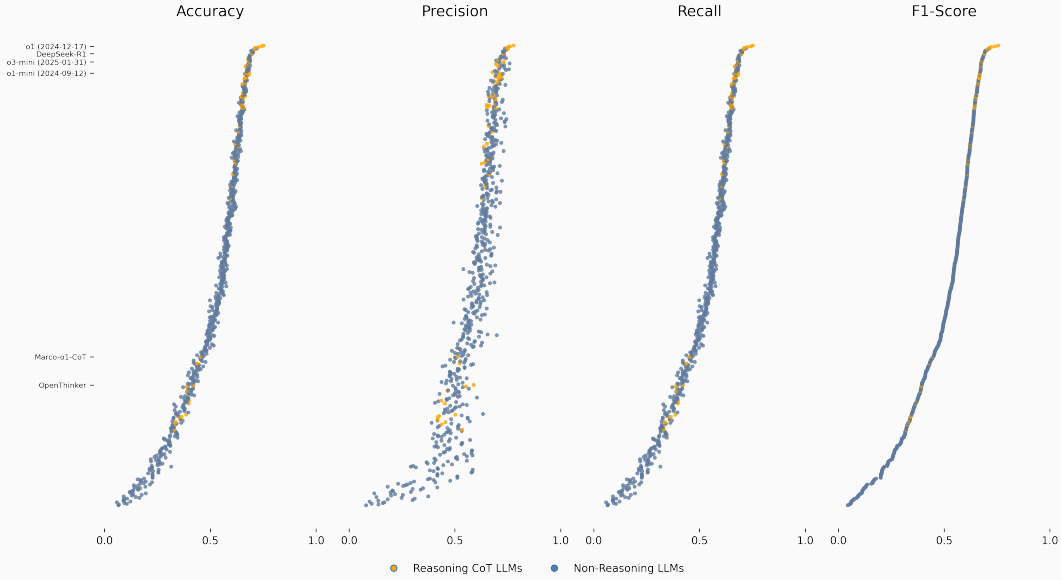
SOTA closed-source LLMs

o4-mini, o3-mini, o1, GPT-4.1, GPT-4.5-preview, Grok 3 Beta, Claude 4 Opus, among others

SOTA open-source LLMs

Llama 4 Maverick (400B) and Scout (107B), Mistral 3.1 (24B), Llama 3.3 (70B), DeepSeek-R1 (671B), DeepSeek-V3 (671B), among others

Benchmarking Policy Agenda


Accuracy
Precision
Recall
F1-Score
o1 (2024-12-17)
DeepSeek-R1
o3-mini (2025-01-31)
o1-mini (2024-09-12)
Marco-o1-CoT
OpenThinker
0.0
0.5
1.0
Reasoning CoT LLMs
Non-Reasoning LLMs

Top Performing Models Policy Agenda

0.76

Maximum F1-Score

Achieved by OpenAI reasoning models

0.69

Top Open-Source Models

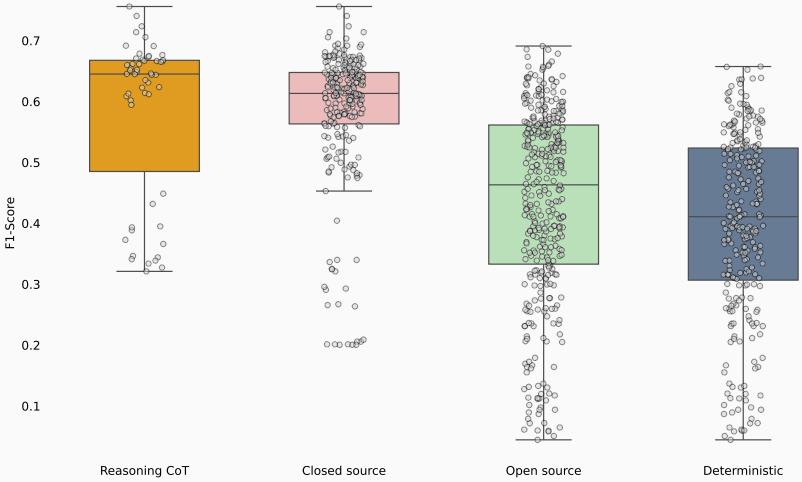
DeepSeek-R1 and V3 (671B) and Llama 3.1 (405B)

0.54

Mid-Range Performance

Gemini 1.5 series, GPT4o-mini, Mistral models, Qwen 2.5

F1-Score Across Reasoning, Source and Deterministic Setup



Comparing with Fine-Tuned Models

Language	Best LLM	F1-Score	Fine-Tuned	F1-Score	Δ Val	Δ Best LLM
Danish	GPT-4.5	0.679	Babel Machine	0.925	+0.065	+0.246
Dutch	o1	0.724	Babel Machine	0.906	+0.066	+0.182
English	o1	0.706	Babel Machine	0.869	−0.031	+0.163
			ft-XLM-RoBERTa	0.810	−0.009	+0.104
French	o1	0.714	Babel Machine	0.821	−0.029	+0.107
Hungarian	GPT.4-5	0.672	Babel Machine	0.751	−0.099	+0.079
Italian	o1	0.675	Babel Machine	0.930	+0.120	+0.255
Portuguese	o1	0.651	Babel Machine	0.867	−0.063	+0.216
Spanish	o4-mini	0.756	Babel Machine	0.916	+0.066	+0.160

Note. All estimates are weighted F1-scores obtained on our fixed held-out test set. The columns Δ Val and Δ Best LLM indicate: (i) the change relative to the best result on the model's own validation set; and (ii) the change relative to the strongest zero-shot LLM, respectively.

How much data leakage is present here? Probably considerable, so that the results may be inflated. However, the potential advantages of fine-tuning still seem to outweigh in-context learning, even for BERT-like models.

Key Findings Meta-Analysis Policy Agenda

$+8.9\%$

Reasoning Advantage

Chain-of-thought capabilities boost F1-score by almost 9 points

-7.8%

Reproducibility Penalty

Deterministic deployment reduces F1-score by about 8 points, but ensures consistency

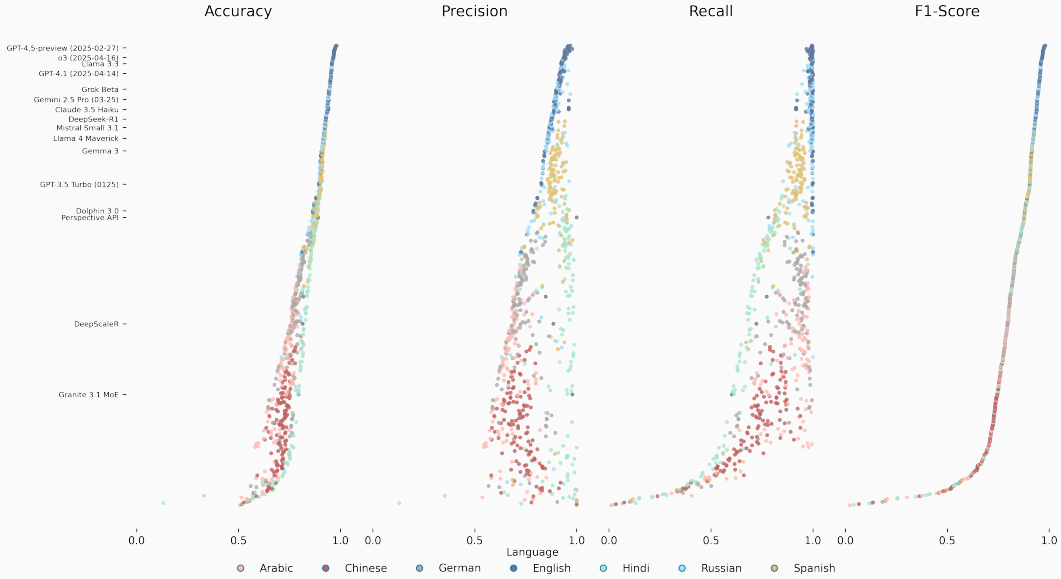
$\geq 10\%$

Fine-Tuning Advantage

Supervised transformers outperform zero-shot LLMs by 10+ points

We did not find 🔒 openness penalty

Benchmarking Toxicity


Accuracy
Precision
Recall
F1-Score
GPT-4.5-preview (2025-02-27)
o3 (2025-04-16)
Llama 3.3
GPT 4.1 (2025-04-14)
Grok Beta
Gemini 2.5 Pro (03-25)
Claude 3.5 Haiku
DeepSeek-R1
Mistral Small 3.1
Llama 4 Maverick
Gemma 3
GPT 3.5 Turbo (0125)
Dolphin 3.0
Perspective API
DeepScaleR
Granite 3.1 MoE
0.0
0.5
1.0
Language
Arabic
Chinese
German
English
Hindi
Russian
Spanish

Top Performing Models

Arabic

0.83 F1-Score

o3 · o1 · GPT-4.5

Chinese

0.78 F1-Score

GPT-4.1 · GPT-4o · Grok 3

English

0.98 F1-Score

Granite 3.2 · OLMo 2 · GPT-4.5

German

0.86 F1-Score

GPT-4.1 · o3 · GPT-4.5

Hindi

0.89 F1-Score

Mistral Saba · Mistral Medium 3 · Gemma 2

Russian

0.96 F1-Score

o3 · GPT-4.1 nano · Claude 4 Opus

Spanish

0.93 F1-Score

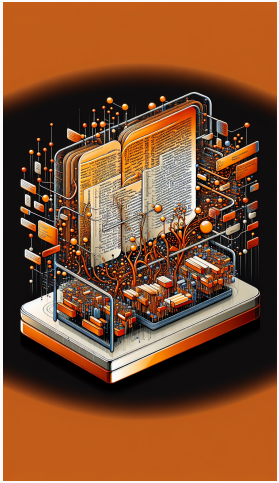
GPT-4.5 · o3 · GPT-4.1

Comparison to Perspective API

Perspective API performance. Competitive accuracy but abysmally low recall scores, resulting in low F1-scores across languages.

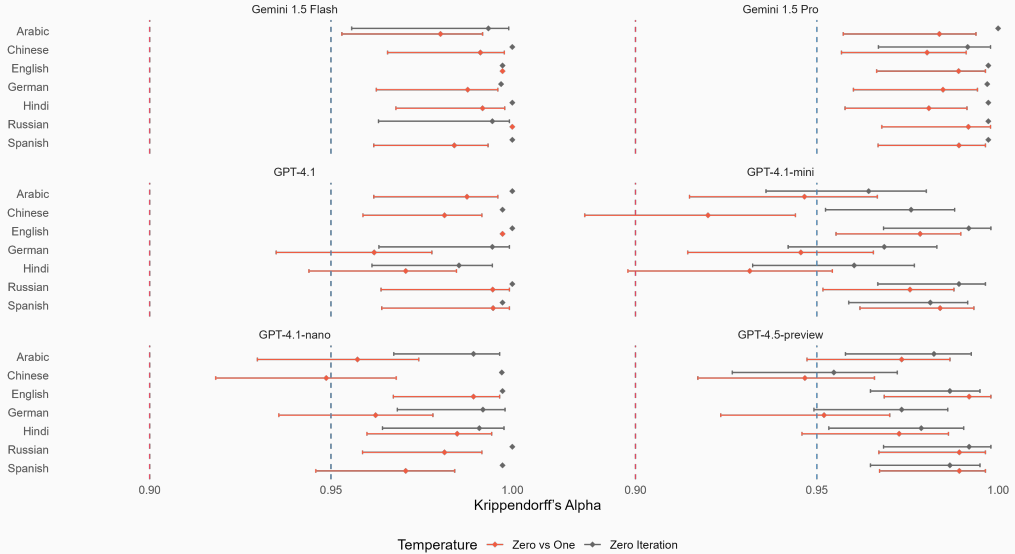
Performance gap. Even weakest LLMs outperform Perspective by about 0.20 points in F1-score for Arabic and Chinese.

Technological shift. Highlights transition from classical transformer pipelines to modern LLMs for toxicity classification.



Artwork by DALL·E 3 model

Temperature Experiments



Key Findings Meta-Analysis Toxicity

+7.7%

Resource Advantage

High-resource languages enjoy a 7.7 points F1 advantage

−9.8%

Size Penalty

Small models ($\leq$7B) fall 9.8 points short of larger LLMs baseline

99%

Annotation Reliability

Modern private LLMs exhibit outstanding reliability even w/temperature variations

We did not find 🔒 openness or 💡 overthinking penalties

Methodological Takeaways

Evidence-Based Recommendations

› **Reasoning advantage.** Chain-of-thought capabilities enhance performance in policy agenda classification, but do not yield gains in toxicity classification.

› **Openness penalty inconclusive.** Slight lower performance than proprietary models, descriptively. There is no significant effect in meta-analyses. However, deterministic deployment reduces F1-scores in policy agenda classification.

› **BERT family pays off in complex tasks.** Supervised transformers outperform zero-shot LLMs by $+10$ points in policy classification. Even the weakest LLMs in toxicity outperformed off-the-shelf options.

› **Performance vs reproducibility.** Switching from deterministic to stochastic decoding rarely alters more than 1% of labels on average in toxicity.

This work has been supported by the Universidad Diego Portales (Inserción/Enlace Fund 2025), Agencia Nacional de Investigación y Desarrollo (ANID/PFCHA/72200340), Society for Latin American Studies (Postgraduate and Postdoctoral Award), St. Hilda's College, University of Oxford (Muriel Wise Fund), and OpenAI (Academic Programme).

Thank you very much!

Do you have any questions?

bastian.gonzalez.b@mail.udp.cl

bgonzalezbustamante.com

Presentation compiled with LaTeX and some

Download the latest version from GitHub